Faculty of Engineering and Information Technology
University of Technology Sydney

Data Mining for High Performance Compression of Genomic Reads and Sequences

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Yuansheng Liu

September 2019

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Yuansheng LIU declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note: Signature removed prior to publication.

Date: 1/10/2019

Acknowledgments

First and foremost, I offer my deepest gratitude to my supervisor, Professor Jinyan Li, for his extensive support and guidance during last three years of my PhD studies. With his help and strong support, I had a chance of winning scholarships. Without his patience, I never would have been uncovered the fascinating areas of bioinformatics. His guidance including scientific writing and research ideas made all of this work possible.

I am very thankful to Dr. Hui Peng, who is our team member and is also my roommate at Xiangtan University. I could not have the opportunity to study at UTS without his strong recommendation. Dr. Peng has vast knowledge in the area of bioinformatics and also gives many insightful suggestions and comments to improve the quality of our projects. I would also like to thank other team members: Chaowang Lan, Tao Tang, Xiaocai Zhang, Xuan Zhang, Yi Zheng and Zhixun Zhao, for their kind help during my studies. And the delicious food cooked by them is really evocative. The fantastic activities with them are enjoyable and relaxing under the heavy pressure of research.

This work would not be possible without the people with whom I collaborated and whose help was essential in many projects I participated in: Prof. Limsoon Wong, Prof. Zuguo Yu, Prof. Quan Zou, Prof. Xiangxiang Zeng and Dr. Yu Zhang.

Furthermore, I gratefully acknowledge all the organizations provided the scholarships including International Research Scholarships, ARC Discovery Scholarship, and the Big Data Big Impact grant.

I am deeply grateful to my wonderful parents and my dear wife for their support and love throughout all my life. I also extend my deep gratefulness to my relatives and friends in China.

As to my lovely daughter and son, my love for them is beyond words. I am very sorry for not being able to accompany in the first two years of them as I study overseas.

Yuansheng Liu

UTS, Sydney, Australia

June 2019

Contents

List of Figures

List of Tables

List of Publications

Below is the list of journal papers associated with my PhD research:

Journal Papers Published

- **Liu, Y.**, Zhang, Y., & Li, J. (2019). Fast detection of maximal exact matches via fixed sampling of query k-mers and Bloom filtering of index k-mers, *Bioinformatics*, in press: 10.1093/bioinformatics/btz273.
- **Liu, Y.**, Yu, Z., Dinger M.E., & Li, J. (2019). Index suffix-prefix overlaps by (w, k)-minimizer to generate long contigs for reads compression. *Bioinformatics*, 35(12), pp.2066–2074.
- **Liu, Y.**, Peng, H., Wong, L., & Li, J. (2017). High-speed and high-ratio referential genome compression. *Bioinformatics*, 33(21), pp.3364–3372.
- **Liu, Y.**, Lan, C., Blumenstein, M., & Li, J. (2017). Bi-level error correction for PacBio long reads. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, in press: 10.1109/TCBB.2-017.2780832.

Abstract

The rapid development of next-generation sequencing (NGS) technologies has revolutionized almost all fields of genetics. However, the massive amount of genomic data produced by NGS presents great challenges to data storage, transmission and analysis. Among various NGS-related big data challenges, in this thesis, we focus on short reads data compression, assembled genome compression and maximal exact matches (MEMs) detection.

First we propose a new *de novo* compression algorithm for short reads data. The method utilizes minimizers to exploit the redundant information presented in reads. Specifically, large k-minimizers are used to group reads and (w, k)-minimizers are used to search suffix-prefix overlap similarity between two contigs. Our experiments show that the proposed method achieves better compression ratio than the existing methods.

Furthermore, we present a high-performance reference-based genome compression algorithm. It is based on a 2-bit encoding scheme and an advanced greedy-matching search on a global hash table. The compression ratio of our method is at least 1.9 times better than the best competing algorithm on its best case, and our compression speed is also at least 2.9 times faster.

Finally we introduce a method to detect all MEMs from pairs of large genomes. The method conducts a fixed k-mer sampling on the query sequence and the index k-mers are filtered from the reference sequence via a Bloom filter. Experiments on large genomes demonstrate that our method is at least 1.8 times faster than the best of the existing algorithms.

Overall, this thesis work has developed efficient algorithms for pattern discovery from and for data compression of genomic sequences of big size.